DOCUMENT RESUME

ED 199 298 TM 810 332

AUTHOR Popham, W. James
TITLE Teacher Evaluation--The Wrong Tests for the Right Job.
PUB DATE Apr 81
NOTE 9p.; Adapted from a symposium presentation at a jointly sponsored session during the Annual Meetings of the National Council of Measurement in Education and the American Educational Research Association (Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; *Norm Referenced Tests; *Teacher Evaluation; *Test Validity

ABSTRACT

Teachers should be evaluated chiefly by the results they produce. Those results will not be properly detected through the use of norm-referenced achievement tests (NRT) due to the following major deficits of NRT: (1) the descriptions of what is measured by NRT are far too loose; (2) evaluative expectations are unclear; (3) the necessity for NRT to produce sufficient variance in examinee responses makes it less likely that such tests will measure pupil mastery of key knowledge and skills; and (4) most NRT do not differentiate among students chiefly on the basis of what those students have learned in school--rather their differentiations are based heavily on the nature of the semantic experience that students bring to school. Pupil test performance should be used as a major factor in any teacher evaluation system, however. Those tests should be well constructed criterion-referenced tests (CRT) which will provide teachers with a clear set of instructional targets and test items which can detect the effects of skillful instruction, communication of evaluative criteria, absence of the automatic excision of items on which pupils perform well, and a vehicle which can measure affective gains. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED199298

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. J. Popham
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TEACHER EVALUATION--THE WRONG TESTS FOR THE RIGHT JOB*

W. James Popham
University of California, Los Angeles
and
The Instructional Objectives Exchange

Teachers ought to be evaluated chiefly on the basis of their pupils' performance. In contrast to the teacher evaluation schemes that focus on process, such as the techniques used by teachers during instruction, I am completely in favor of teacher evaluation systems that emphasize produce, such as who much students have learned or what sorts of attitudes they have acquired.

Perhaps my unbridled distress with process-oriented teacher evaluation systems stems from their almost cavalier discounting of what happens to pupils as a consequence of the teacher's efforts. I don't care if Mr. Smith can lecture with the eloquence of Demosthenes; if his students don't acquire worthwhile skills, knowledge, or attitudes, then he's a lousy teacher. I don't care if Mrs. Jones can conduct group discussions with suave that would induce envy in Carl Rogers; if good effects don't occur for her students, then she's a rotten teacher. It's what happens to students as a consequence of the teacher's procedures that should serve as the major index of the teacher's prowess. One of the best ways of gauging instructional effects, but by no means the only one, calls for the use of pupil test performance in the appraisal of teachers.

TM 810 332

Now there are obviously all sorts of difficulties associated with the implementation of teacher appraisal schemes based on the test performance of pupils. Not the least of these is the complaint, often voiced by officials of teachers organizations, that the adverse effects of certain parental or subcultural environments are so powerful as to render ineffectual the efforts of even superlative teachers. But while this sort of problem is well recognized, there is another difficulty in using test results to evaluate teachers that is less well known, namely the use of achievement tests that are fundamentally inappropriate for that purpose.

Teachers who allow themselves to be judged on the basis of student performance on traditional, norm-referenced achievement tests are participating in an essentially unsound, hence, unjust, enterprise. It is often true, however, that teachers do not understand why it is that the use of norm-referenced tests in a teacher appraisal scheme constitutes such an egregious mistake. This analysis will, in brief, provide an explanation of two major shortcomings of traditional tests when used in a teacher evaluation context.

Mushy Targets

One of the major deficits of norm-referenced achievement tests is that the descriptions of what they actually measure is far too loose. Because norm-referenced tests function chiefly to permit comparisons among examinees, typically with respect to a fairly broad set of knowledge or skills, it is not really requisite for the creators of these tests to build descriptions which are particularly precise. There is no pressing need to construct super-detailed test specifications that delimit the range of permissible test items. There is no pressing need

to make certain that all items are congruent with sets of detailed test specifications. If the test serves its purpose of producing sufficient response variance among examinees so that comparisons can be made among them, then terribly precise descriptive information regarding what's being tested is not needed.

Most norm-referenced achievement tests, therefore, are accompanied by completely inadequate descriptions of what is being assessed. There may be a few vague generalizations, or perhaps a list of content topics, but little else to tie down the classes of behavior supposedly being measured. Sometimes a norm-referenced test's technical manual will even include a "table of specifications." Close analysis of those specifications reveals a degree of descriptive rigor which is, at best, murky.

The difficulty, from the teacher's perspective, is that the evaluative expectations are unclear. Thus, we find teachers engaging in a guessing game which ends up sounding much like the pre-exam "What'll" queries of their students, that is, "What'll be on the exam?" "What'll he ask?" or "What'll she emphasize?" And, just as their students, teachers often guess wrong about the What'll answers.

Given the diffuseness of most descriptive information in norm-referenced tests, teachers typically fail to get a solid fix on what's to be measured. And, without a clear notion of what's to be tested, it is next-to-impossible for teachers to provide a truly on-target instructional experience for their students which includes sufficient time-on-task practice opportunities for those students. After all, the teachers don't know what the task is. To evaluate professionals on the basis of ill-defined criteria is fundamentally unjust.

Psychometric Subtleties

The second major problem with norm-referenced achievement tests, for purposes of teacher evaluation, stems from the psychometric traditions governing the creation of such tests. Central among these traditional imperatives is the necessity for the norm-referenced test to produce sufficient variance in examinee responses so that fine-grained comparisons can be made among those examinees. These comparisons are made, of course, in relationship to the performance of the original norm group. Test items that contribute most to the production of response variance are those which are answered correctly by about half the examinees. If too few or too many examinees answer an item correctly, then it fails to do its share in spreading out examinee performance. Such items are modified or removed from the test.

But, distressingly, test items on which students perform well are often those dealing with topics that teachers have stressed. Putting it another way, the more important a topic is--the more the teacher will emphasize it. But, the more that teachers emphasize particular topics--the better that students will perform on items measuring those emphasized topics. As a consequence, the better that students perform on test items--the less likely it is that those items will remain in the next version of the test.

America's teachers thus find themselves in a Catch-22 dilemma wherein the more effectively they promote pupil mastery of key knowledge and skills, the less likely it will be that norm-referenced tests will include test items measuring such well taught skills and knowledge. This is particularly true with oft-revised norm-referenced tests, in which high student performance on test items leads to the instant demise

of such items. It is even true, however, with brand new norm-referenced tests since the test's designers will eschew items that are "too easy." From the poor teacher's perspective, phrases synonymous to "too easy" might well be "too often emphasized" or "too well taught."

Test developers frequently resort to the use of verbal nuance in the creation of test items that do an adequate job of discriminating among examinees. The more verbally facile that examinees are, the better they perform on such examinations.

In short, a host of hallowed psychometric practices lead to the creation of norm-referenced achievement tests which are, for all practical purposes, largely incapable of detecting the effects of instruction--either effective or ineffective. Most norm-referenced achievement tests do not differentiate among students chiefly on the basis of what those students have learned in school; rather, their differentiations are based heavily on the nature of the semantic experience that students bring to school.

Pupils whose subcultures force them to experience, day in and day out, the set of verbal subtleties that test constructors employ to differentiate among examinees are at a tremendous advantage when it comes to taking traditional norm-referenced achievement tests. Pupils whose subcultures are largely devoid of those verbal subtleties will encounter substantial difficulties when taking achievement examinations constructed according to traditional norm-referenced tenets.

I do not mean to suggest that norm-referenced achievement tests measure only what the child brings to school. Scrutiny of such tests will reveal that there are scores of items that a totally un-schooled child would have no chance of answering. I am contending, however, that most norm-referenced tests base a major segment of their discriminating

efficiency on factors largely unrelated to the effects of school-based instruction.

To the extent that this contention is accurate, we see once more an instance of educational injustice. Teachers are being evaluated on the basis of testing devices which have been constructed in such a way as to render teacher impact minimal.

Right Job--Right Tests

I am thoroughly in support of the use of pupil test performance as a major factor in any teacher evaluation system. But those tests should be well constructed criterion-referenced tests, not norm-referenced measures.[1] Properly developed criterion-referenced tests will offer teachers a clear set of instructional targets and test items which can detect the effects of skillful instruction. Properly developed criterion-referenced tests will provide teachers with the evidence needed to create a truly equitable method of teacher evaluation.

Well formed criterion-referenced tests communicate with little ambiguity what the nature of the evaluative criteria will be. Teachers have a reasonable chance to create an on-target and potent instructional sequence because they know what the targets are. They can provide ample practice opportunities for learners so that they master the class of skills being measured. Besides that, if they're astute, teachers can describe the nature of the intended targets to their pupils, thus clarifying for their pupils what is expected of them.

Well formed criterion-referenced tests do not automatically excise items on which pupils perform well. If the items are consonant with the skill being measured, as represented in the test's specifications, then a test item stays in the test when it's revised even if most pupils

answer that item correctly. Teachers who promote learner mastery of important skills and knowledge need not worry that the key items measuring those skills and knowledge have been chased from the test.

Well formed criterion-referenced tests do not need to be only cognitive in nature. Some of the most important consequences that teachers promote are in the affective domain. Affective assessment devices can be created along criterion-referenced lines, then incorporated into a meaningful teacher appraisal system.

In sum, I believe that teachers should be evaluated chiefly by the results they produce. Those results will not be properly detected through the use of norm-referenced achievement tests. Teachers should rebel against the use of such tests in any teacher appraisal system. In contrast, the use of criterion-referenced tests offers teachers the opportunity to be evaluated by means of a genuinely just teacher evaluation enterprise.

Footnotes

*Adapted from a symposium presentation at a jointly sponsored session during the annual meetings of the National Council of Measurement in Education and the American Educational Research Association, Los Angeles, April 13-17, 1981.

[1]For a description of the elements comprising a well constructed criterion-referenced test (as well as everything else you've ever wanted to know about measurement), see Popham, W. James, Modern Educational Measurement, Prentice-Hall, Inc.: Englewood Cliffs, N.J., 1981.